

## Methodological notes

P7 Data Scientist OpenClassrooms

Started on Monday the 28<sup>th</sup> of November 2022

Oumeima EL GHARBI

## La méthodologie d'entraînement du modèle

(2 pages maximum)

- Feature engineering
- GridsearchCV
- Cross Validation (number of folds)
- Training score
- SMOTE
- Metrics : AUC, recall, F-score

# La fonction coût métier, l'algorithme d'optimisation et la métrique d'évaluation

(1 page maximum)

## Evaluation metrics

### Use case metrics

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

- In our case, predicting 0 (Negative) means that the client will repay the loan.
- Predicting 1 (Positive) means that the client will not repay the loan (**default risk**).
- We have an imbalanced dataset : our database of clients for which we know if they have repaid their loan contains 92% of clients that have repaid it while 8% had a default of payment.

## ROC Curve and AUC

For the Kaggle competition, AUC-ROC score was chosen to compare the different model performances.

**True Positive Rate (TPR)** is a synonym for recall and is therefore defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

**False Positive Rate (FPR)** is defined as follows:

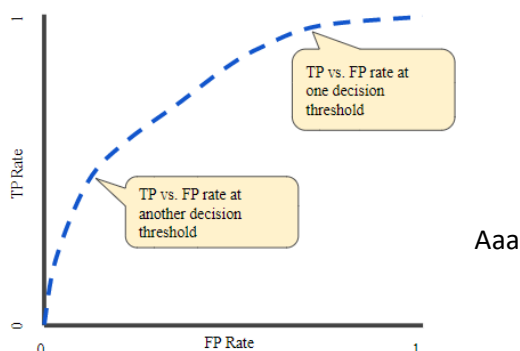
$$FPR = \frac{FP}{FP + TN}$$

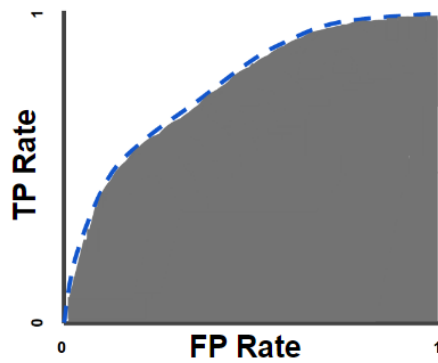
Actual	0	1	
	TN	FP	
1	FN	TP	
		Predicted	
		0	1

False Positive Rate (FPR) =  $\frac{FP}{FP + TN}$

True Positive Rate (TPR) =  $\frac{TP}{TP + FN}$

## ROC curve





Google : "AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1)."

<https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>

## Precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

- In our case :
- FP (False Positive) means that our model predicted 1 when it should be 0.
- If we have False Positive, it means that our model predicted that the client will not repay the loan when in fact the client might repay it.
- This means the bank (Home Credit) lost one potential client that could have repaid the loan without trouble.
- Thus, if we have a high precision, then we won't miss potential client that could have repaid the loan. A low precision means we will miss "good" clients.
- However, the banker can decide to give a loan by studying the client's application more in detail.

## Recall

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

- FN (False Negative) means that our model predicted 0 while it is positive.
- **This means that our model predicted that the client will repay the loan while in fact the client will not repay it !**
- This is important for our bank. Indeed, we do not want to have a high number of False Negative.
- Having a high number of False Negative means that Home Credit gave a loan to clients that might not repay it !
- Therefore, we want to reduce as much as possible giving loans to clients that might not pay it back, thus, we need to lower the number of False Negative which means we want to get a good recall.

**Conclusion** : precision / **recall** : we prefer having a good recall instead of a good precision. This means that we want to lower the number of False Negative which means that we do not want to give a loan to clients that might not pay it back.

## Loss / Cost function

### F-score

*F1-score (beta = 1)* : recall as important as precision

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}.$$

### *F Beta score*

A more general F score,  $F_\beta$ , that uses a positive real factor  $\beta$ , where  $\beta$  is chosen such that recall is considered  $\beta$  times as important as precision :

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}.$$

**We have to choose a value for Beta with  $\beta > 0$ .**

- **As for the F-score, we need to give more importance to the recall for our use case.**

For example,  $\beta = 3$  means that recall is 3 times more important than precision.

*From sklearn documentation :*

*“The beta parameter determines the weight of recall in the combined score.  $\beta < 1$  lends more weight to precision, while  $\beta > 1$  favors recall.*

*For the limits, when  $\beta \rightarrow 0$ , we consider only precision,*

*If  $\beta \rightarrow +\infty$  only recall).”*

## L'interprétabilité globale et locale du modèle (1 page maximum)

### Shapley value

Global

aaa

Local

aaa